Webinar: Determining the Privacy-loss Budget
Research into Alternatives to Differential Privacy

June 4, 2021

Michael Hawes:
Thank you, operator. And good afternoon everyone and thank you for joining us for the latest in our Webinar series on understanding the 2020 Census Disclosure Avoidance System. This month the Census Bureau's Data Stewardship Executive Policy committee or DSEP will be making determinations on the Privacy-loss Budgets and system parameters to be used for the production run of the Public Law 94-171 redistricting data summary file.

Now we've gotten a lot of questions about how DSEP will be making this decision and what information they will be considering in their deliberations. Most of the inputs of that decision making will be a variety of accuracy metrics that we've developed in consultation with our stakeholders as well as the feedback and analyses that we've been receiving based on the April 2021 demonstration data that we released.

But to properly evaluate the privacy and accuracy trade-off that they're going to have to make it's also really helpful to consider what the alternatives would look like were we to use a different form of disclosure avoidance for the 2020 census. So, to contextualize that we have conducted over the past year-and-a-half we've conducted me extensive research on what the privacy and accuracy consequences of different disclosure avoidance methods would be. So, the presentation today will examine the results of some of that research into alternatives to differential privacy that DSEP will be considering as they make their determinations about the system parameters to be used for the production run of the redistricting data.

Before I start I want to thank my many colleagues at the Census Bureau and our external partners who have contributed to much of the information that we've been sharing throughout

this webinar series. And I'd also especially like to thank our many external stakeholder groups who have been providing invaluable feedback throughout the entire process that has helped us to improve our Disclosure Avoidance System over the past couple of years. I also want to state that any opinions and viewpoints that are expressed today are entirely the opinions of the speakers and do not represent the opinions or viewpoints of the US Census Bureau.

Now throughout the presentation you're welcome to submit questions. Please use the Q&A feature in the Web platform. And my colleagues Jen Shopkorn, Meghan Maury, and Michele Hedrick and I will be answering some of those during during the presentation. And then we'll have time at the end to answer any remaining questions that you might have.

So, jumping in the selection and implementation of any disclosure avoidance method imposes a fundamental trade-off between privacy and accuracy. These impacts on the usability of data aren't a by-product of confidentiality protection they're the mechanism by which you protect privacy and confidentiality.

And you can think of this as a spectrum between one extreme where you have perfect privacy with useless data and the other extreme where you have perfect data with no privacy protections at all. And ultimately you need to find a point in between those two extremes where the data are going to be sufficiently protected while also being sufficiently accurate. Once you've identified that point between those extremes that becomes known as your Privacy-loss Budget -- and you'll hear us refer to that as PLB for Privacy-loss Budget -- are represented by the Greek letters epsilon or rho depending on the context.

Now determining that ideal point in between those extremes where the data are sufficiently protected and sufficiently accurate is a policy decision. And it's a difficult and challenging policy decision at that because it requires weighing kind of countervailing responsibilities that we have as an agency to produce high quality data while also meeting our legal and ethical obligations to protect the privacy of our respondents and the confidentiality of their data.

Now in making that determination of where on that spectrum to be you have to use a lot of different factors. As I mentioned before DSEP will be considering a broad and diverse array of accuracy measures that we've developed both those included in the detailed summary metrics that we published with our April 2021 demonstration data but also additional accuracy metrics

that we've developed for specific people in the Redistricting and Voting Rights Act use cases as well as accuracy targets related to many of our demographic programs that use census data including our Population Estimates Program.

But in addition to evaluating all of those accuracy measures we'll also be evaluating privacy impacts, both in terms of the privacy guarantee that could be afforded by various levels of Privacy-loss Budget and by the comparisons that we can make to similar privacy protections with different types of disclosure avoidance applied. So, the point of today's presentation as I mentioned before is to look at okay, what will the impact on privacy and what will be impact on accuracy and availability of data be if we were using a different disclosure avoidance method in place of the 2020 Disclosure Avoidance System that we have. And being able to make those comparisons allows us to contextualize those trade-offs that we have to make in the subsequent days and weeks.

So, with that I'm going to turn things over to my colleague Rolando Rodríguez who is going to walk through the results of the research that our teams have done on this subject. Rolando?

Rolando Rodriguez:
Thank you Michael and good afternoon everyone. So, I'll start off with a bit of background on the DAS Reconstruction Team and the effort that team has made over the past year and a half.

Next slide. We skipped one Michael. So, the DAS Reconstruction Team is a separate team from either the DAS Science or Dev Ops teams. Those teams are working steadily to finalize the implementation of the top down algorithm for use in 2020 Census production as a means of providing privacy to our respondents.

So, back in February of last year a group in the Center for Enterprise Dissemination Disclosure Avoidance area was assembled to begin to assess the potential impact of swapping using an algorithm based upon the 2010 swapping algorithm. That team has since become the DAS Reconstruction Team. We perform numerous of these swapping experiments and we have also begun preliminary assessments of the potential impact of cell and table suppression on 2010 data.

Next slide. So, I'll start out talking about suppression and the tests we have done there. These

experiments involve 1980 Census suppression rules which was the last census we actually used suppression as the main means of disclosure avoidance. And these are based upon OMB race categories that I'll describe in a second.

Next slide. So, what is suppression? So, suppression involves the removal of information from published tables in order to help protect privacy. In the 1980 Census there were two types of suppression used: table suppression so the whole removal of tables from publication, and cell suppression where particular cells and certain tables are suppressed to provide privacy.

Typically, cell suppression is the harder type of suppression to implement due to the need for what we call complementary cell suppression which I'll explain here. Next slide. So, when you do suppression you have some kind of rule. So here we have a two way table and its associated margins that are going to be released.

You may have rules that say any cell that has fewer than five, a count of five associated with it will be suppressed. So, in this table we would take the two here and replace it with a suppression value here in S. In the 1980 Census they used a zero. So the thing is we're not done here. This is what we call only a primary suppression.

Next slide. Turns out that primary suppressions aren't enough to provide the kinds of protection that you want from a suppression method. And so in this case since we are publishing the margins of this table you can take the margins and do just simple arithmetic to get back to the value of S. So, what do you have to do you have to make what are called complementary suppressions to support the actual suppression of the cell you want. In this case because we have a two way table and we have all the margins published you actually have to end up suppressing the entire set of two way cells.

Next slide. So, for the suppression test the DAS Reconstruction Team assessed the impact of 1980-based suppression rules. So, the counterfactual world here is that we have suppression rules based upon the 1980 Census. We have race and ethnicity categories given by OMB. And we're applying these to the PL94-171 redistricting tables and the Summary 1 File tables from the 2010 Census. And these are based on the 2010 Census Edited File.

So, the race categories I've looked at here as I said are given by OMB and implemented by the

Department of Justice Voting section. There are seven major race categories either alone or in combination with white but the remainder are two or more race category. And then for ethnicity we have the simple Hispanic or non-Hispanic binary classification.

Next slide. So, as far as the rules for the 1980 Census and suppressions there were a couple of rules that were germane to the PL94 tables. For table suppression the rule was that whole tables were suppressed, that is weren't published at all, or any geography that had between one and 14 persons in any of the race or race by ethnicity groups.

So, this applied to two tables, table P3 which is race for the voting age population, and table P4, which is race by Hispanic or Latino origin for the voting age population. The other rule that applies is cell suppression of counts of one or two. Those were suppressed and marked as zero. There was a flag to indicate which zeros were actual sampling zeros and those that were zeros from suppression. That applied to two tables, tables P1 which is race alone, and table P2 which is ethnicity by race.

Additionally for the SF1 tables from the 2010 we looked at table suppression rules from 1980 that said that any table that wasn't dedicated solely to race or ethnicity would be suppressed if the geography associated with that table had between one in 14 persons. Through a simple suppression based upon total population that applied to all person level tables from SF1.

Next slide. So, I'm going to talk a bit about the impact that these suppression rules would have on privacy risk. So, with suppression if it's done correctly you actually do remove information from the tables that are released. If you do enough of this suppression you can prevent particular attacks for instance like the reconstruction-abetted re-identification attack that was performed on the 2010 Census. And while we can eliminate the risk of a specific attack this wouldn't be equivalent to the kinds of broad privacy protections that are associated with formal privacy definitions like the one being used for the 2020 TopDown Algorithm.

Next slide. So, to get some actual results here I have a summary of results. This is cell suppression for tables P1 and P2 out of the PL94-171 products for 2010. Again this is using race categories from OMB. These are only the primary suppressions for the cells that I'm showing.

I'll note that the margin of these two tables, P1 race and P2 Hispanic by race, to the total the

population margin there was always reported under the 1980 rules. So, again you would need complementary suppression above these numbers.

So, what we see here is that at the county level for table P1, 2.4% of the cells will require suppression, and 10.1% of the cells at the block group level. Table P2 since is more - since it crosses table P1 by another variable it will have more cell suppression that would need to be applied. At the county 6.8% of cells up to a maximum of 14.5% of cells at the block group level.

Next slide. More interesting perhaps is the effect of table suppression. So, tables P3 which is race for the voting age population and table P4 which is race by ethnicity for the voting age population. Since those cross race or ethnicity by another demographic, in this case age, they would have had table suppression applied in 1980.

So, we show here some of the results for potential suppression. Under these rules we see for table P3 at the county level almost over 50% of the tables would have to be suppressed, reaching a maximum of 95.7% of tables at the block group level.

Again, for table P4, since it crosses the previous table with another variable, in this case it's now race by ethnicity, it would be even a larger amount of table suppression would have been necessary. At the county level for table P4, 84.2% of tables would have been suppressed, of the P4 tables that is, reaching basically almost all the tables at the block group level would have needed suppression. So, this is a large amount of suppression that would have been necessary under these rules especially for on-spine geographies lower than the county level.

Next slide. There was not a cell suppression rule for tables P3 and P4 in 1980 but we did consider if there were, what would have been what in the implications. So, again here we have - this would be primary cell suppressions. And similar story to tables P1 and P2. For table P3 at the county level 3.7% of cells would be suppressed up to a maximum of 13.5% at the block group level. Table P4, again that's Hispanic or Latino by race for the voting age population over 9% of cells at the county level would be primary suppression up to a maximum of 17.6% at the block group level.

Next slide. But as I said before we additionally looked at rules would have affected other SF1 tables. So, these are tables including finer breakdowns of race - of sex and age. We looked at

the potential impact on both the person level tables and the housing unit level tables that roll to housing levels tables with that. If you had between one and four occupied housing units in a geography the SF1 table for that geography would be suppressed.

So, as we see here so again this is table suppression for SF1 tables based upon the 1980 suppression rules. For blocks of 38.7% of the SF1 - of the blocks we need to have the SF1 table suppressed. For the housing units as a block again 32.8% of blocks would have to have their housing unit SF1 table suppressed under these rules. And you can see here too that there are additionally some need for a little bit of table suppression at - of block group tables and tract level SF1 tables as well.

Next slide. So, those are the results related to suppression I'll now move on to our results related to swapping. These are relaxations and extensions of the swapping algorithm that was used for the 2010 Census.

Next slide. So, just for the overview of how swapping works you first have to determine a key that's going to look to match units. Those units will be the swap partners with each other. So, they're matched on a certain key.

You have to choose which geographies you're going to swap between and that you'll swap within. So, as I have illustrated here we're using a matching key of the number of people in a household. We're swapping between blocks and but within say tract, county or state.

You then determine which units you want to swap. That then feeds into the calculation of the swap rate that you want. So, what percentage of the housing is in this case would be swapped. And then you find pairs. And if you find a pair you swap the pair. So, this moves households between geographies and in that way tries to provide some amount of privacy.

Next slide. So, initial efforts from the DAS Reconstruction Team focused on taking the 2010 swapping algorithm and making it support high swap rates up to 100% if necessary. So, they made those changes so the algorithm as we have it now accepts a few parameters. One is that the desired swap rates are now anywhere between zero and 100%.

I looked at some variance of the key but you want to be constant between the swap - the

household that you're swapping. And then we have mechanisms for relaxing those invariants and extending their swaps beyond the tracts because often swaps end up happening within tract under this algorithm.

Next slide. So, we've prepared numerous iterations of these swapping experiments at this point. We have tested swaps race ranging from 5% to 50% of housing units. We've applied pre-swap perturbation to household size of up to plus one or minus one for up to 80% of housing units.

We've applied pre-swap pertubation of tracts either within county or within state to housing - for up to 70% of the housing units. Then at the beginning of this year we began to set the impact of these parameters have on the reconstruction-abetted re-identification attack that has been previously done on the 2010 Census.

Next slide. So, what are the key outcomes of these swapping experiments? Well one is that swap rate has had essentially no impact on our re-identification outcomes. They stay essentially the same as they were for the 2010, original 2010 SF1 experiment.

High rates had a minimal impact on the re-identification outcomes but had accuracy metrics significantly inferior to those out of the latest DAS PPMF test released in April. Put those together that implies that there's some middling swap rate where you might be able to match the TopDown Algorithm in terms of accuracy and particular outputs. But we're going to have low impact from swapping on reducing re-identification risk at least in these scenarios.

So, to give some particular results here as to recall what happened for the 2010 original experiment we had a (putative population) re-identification rate of 44.6%, a confirmed percent across the population of 16.85%. That gives us a precision. So, that is the relative amount of confirmation to the putative re-identification of 37.79%.

But after a two swap experiment, one I'm calling "SwapLow" and one I'm calling "Swap High," SwapLow is a 5% swap with no perturbation of household size or tract. SwapHigh was a 50% swap with 50% of the households having their household size perturbed pre-swap and 70% of them having their tract perturbed pre-swap. For the SwapLow experiment you can see the putative and confirm conversion rates basically mirror those from the original experiment,

meaning the precision also stays the same.

For the SwapHigh experiment, you know, a definite decrease in putative and confirmation rates. But that's only a putative rate down to 42.69% and a confirmation rate and the population down to around 13%, which means our precision still stays above 30%.

Next slide. So, then looking at some of the impact on accuracy of these swap experiments I have here mean absolute error for the total population in the county or total population for incorporated places. You can see here I have the April PPMF release the SwapHigh and the SwapLow experiment. You can see that the errors from the swap high experiment dwarfed those from either of the others. And again that had some but pretty minimal impact on re-identification outcomes.

Next slide. We also broke this down for race alone categories at the county level. So, this again mean absolute error. We see the same story again. SwapLow in the April PPMF are basically both invisible on the scale. The errors from swapping 50% of housing units it's just incredibly high for these kinds of outcomes.

Next slide. So, just some final considerations to take away. None of these algorithms either for suppression or the swapping algorithms adheres to a formal privacy definition of semantic for privacy loss like the ones that are being used for the 2020 top down algorithm. We are only assessing them on this one particular attack strategy the 2010 Census reconstruction-abetted re-identification attack.

The Census Scientific Advisory Committee recently stated of course that this was a conservative attack. And even using just our own census data we can do much more specific attack than this. The implementation of the 1980 Census suppression rules would lead to extreme amounts of table suppression for the sub-state on-spine geographies, So, county, track block, and block group.

If we implement these relaxations or extensions of the 2010 Census swapping algorithm that would yield little improvement in our re-identification outcomes, at least for the particular experiments we did on 2010. And the accuracy outcomes at the highest swap rates would be significantly inferior to those out of the current DAS.

As a final note here, any production implementation of either a suppression or swapping algorithm would be expected to take at least an additional six months after a decision would be made to implement them. Next slide.

Michael Hawes:
Great, thank you Rolando. Before we open things up for some of the questions that have been coming in during the Q&A I just wanted to let everyone know if you're interested in staying informed about our work on the disclosure avoidance system and our modernization of disclosure avoidance for the 2020 Census and for our other censuses and surveys please sign up for our newsletter. You can find that on census.gov if you search "disclosure avoidance."

Also check out our Web page. We have lots of great information there. We have issue papers, frequently asked questions, fact sheets, videos and much, much more. If you want to learn more about what the work is that we've been doing and stay tuned to the newest updates. So, with that I'm going to turn things over to Meghan Maury who is going to moderate our Q&A. Meghan, take it away.

Meghan Maury:
Thank you so much, Michael. And actually this first question is for you even though I see that you were just moved off the stage. A couple folks have asked about something that you talked about in the intro which is they were asking about the decision-making process that's happening right now around the epsilon and allocation of the PLB. Can you just revisit that and maybe talk a little bit about what happens, how that whole process works?

Michael Hawes:
Sure, sure. So, as I mentioned, DSEP is expected to make their determinations for Privacy-loss Budget and system parameters this month. And that decision-making is is currently underway. Once DSEP has decided on the Privacy-loss Budget the allocation of the Privacy-loss Budget and the other system parameters for the production run of the redistricting data product those specifications will be given to the disclosure avoidance system team for implementation in the production environment that will be used.

Then we will be doing the actual production run of the redistricting data. The output of that is

the Microdata Detail File. That Microdata Detail File will then go through extensive quality assurance evaluation by our demographic directorate. And that essentially serves several purposes but primarily it's to make sure that the implementation of the parameter settings that DSEP selected were properly implemented and didn't result in any errors or unexpected consequences.

So, there's this QA period in there to make sure that the data are behaving as expected. At that point they then fed into tabulation. They'll be tabulated and then that moves towards the August 16 release of the redistricting data product in the legacy file format and the September 30 release of the redistricting data in the data.census.gov platform.

Parallel to those final stages of the process the Disclosure Avoidance System team will also be generating a new set of demonstration data based on the 2010 Census run through the algorithm with the final parameters and settings as implemented in the MDF that becomes the official 2020 redistricting data product.

And so we'll be releasing that additional set of demonstration data using the 2010 data with the final 2020 production settings in September. And that will allow data users to evaluate the anticipated accuracy of the actual 2020 redistricting data products.

Meghan Maury:
Awesome. And one follow-up question to that before I come over to you Rolando with a bunch of fun questions about suppression and swapping. The one follow-up that I saw in the Q&A was, "Can you say what you mean by parameters that DSEP is considering?"

Michael Hawes:
Sure. So, for those of you who attended our Webinar on the mechanics of the top down algorithm: the algorithm is incredibly flexible. And in fact it has nearly infinite tuning parameters that can be used to prioritize accuracy and really like tune in on specific accuracy targets for any given statistic or to tune in for privacy protections for any particular statistic or attribute.

And so there's lots of these different dials or settings that have to be made within the system. The most obvious is of course the overall Privacy-loss Budget to use. But then there's also the

Privacy-loss Budget allocation. How much do you assign to tabulations at the national level versus the states, tract, block group or block levels?

How much do you assign to the different queries that support the production of the data? So, how much do you assign to queries on total population versus queries on race or race by Hispanic origin? How much do you put towards the full detailed table, which is every variable by every other variable? And those ratios determine the relative accuracy of the different sets of tabulations at the different levels of geography.

Then you also have additional parameters that we've talked about in prior Webinars including aspects of the geographic hierarchy, the actual query strategies that are used to support the process in the DAS and much, much more. So, when I say the system parameters that's kind of shorthand for many of those individual little dials and settings that can have implications on that fine-tuning of the algorithm to meet a very precisely kind of calibrated privacy-accuracy trade-off.

Meghan Maury:
Thank you, very helpful. And Rolando this one's for you and it's my favorite question which is, What does perturbed mean in the context of these swapping studies?

Rolando Rodriguez:
Right. So, when we say we perturbed housing units by pre-swap what that means is that we gave some probability that any given housing unit will have its value of size changed by plus or minus one in this case. So, you know, if we did 80% of housing units that get perturbed that means we have some split from about 80, say, 40/40 where 40% go to plus one and 40% go to minus one. So that means that they are matching based upon a housing unit size that is not their own originally.

For geographic perturbation we take pre-swap, we give some probability that the tracts associated or, you know, blocks, or block or tract whichever you want, that's associated with the housing unit gets changed to another geography within the geography that we're doing. So, in this case if we're perturbing tracts we have some probability that a given housing unit will change them to have a tract that's different from the one that had originally which means that they can now swap - they will swap based upon that what we call psuedo tract. So, there it's

actually possible they might go back to their original tract but there is a perturbation.

Meghan Maury:
Got it. And that connects into another question that I saw in the Q&A, which is, they asking that question, "Didn't swapping keep the count invariant at the block group or block or block or track level?" And I think the answer is you tested the perturbation at different levels of geography to see how that would impact these analyses was that right?

Rolando Rodriguez:
Perfect. Well one of the analysis we did had no perturbation. So that would hold. In this case the only thing that would be maintained would be the housing units size between the housing units. So, but, you know, we did have to maintain for instance, like voting age population I believe in 2010. So they were definitely invariants that were required of those swapping algorithms just as for the TopDown algorithm in 2020.

Meghan Maury:
Got it. And then I think a pretty connected question here is that, "In this presentation you provided results about mean average error for county and place. Is there anything that you could tell us about the accuracy of the tract, block group or even block levels for those?"

Rolando Rodriguez:
Yes. And so, like Michael said, you know, the main reason for this was to help inform the staff in terms of their decision on Privacy-loss Budget. So, you know, we have - when we run the swaps we run the entire set of accuracy metrics that are run for the 2020 DAS experiments. But we haven't released any of those other than ones we have here publicly currently.

Meghan Maury:
Got it. There's a question in the Q&A about there's so many in here I'm trying to pull ones that we're getting multiple times. But there's a question here about, "Did you test the re-identification and reconstruction risk for the 1980 suppression rules or was there another way you were analyzing risk of re-identification for ,those experiments?"

Rolando Rodriguez:
Right. So, to understand what the suppression experiments we didn't actually perform

suppression we assessed how often the rule will be applied. So, in fact one of the difficulties and why these kinds of things would take at least six months, is that if we implemented suppression we would have to actually do suppression which means we would need software to perform all the complementary suppressions. And that's a hard task so no, we were not able to perform the same kind of analysis on the suppression tables.

Meghan Maury:
Got it. There are a couple of questions in here that are really legal questions. And I know that neither of you are attorneys so I'm going to put those to the side for now and say that if folks do have questions about legal matters feel free to email us. We'll answer it if we can. As I'm sure that most of the folks on the call know we're not able to speak to current litigation matters. So, to the extent that you're asking a question that's directly related to that it's a little tough for us to answer. But we will answer if we can. And certainly we'll provide you links to information that's on our Web site that might help answer your legal questions on the Census Bureau - on the laws and statutes that guide our work.

A few - many other questions in here. Let me just try to get some easy ones out of the way. One of the questions is, "Have the current demonstration data products, the PPMF that we've been released, have those been reviewed by our demographic directorate? And if so, can we see the results of those analyses in some way?"

Michael Hawes:
So yes, not just the data that have been released as demonstration data files but also many, many of the other experimental runs of Disclosure Avoidance System TopDown Algorithm that we've been performing over the last many months. In fact the demographic directorate and particularly the population division have been enormously active in the ongoing discussions and tuning and tweaking of the algorithms designed throughout the process. And we would not be where we are without their subject matter expertise.

The majority of the analysis that demographic directorate has done has focused on the various measures that are reflected in the detailed summary metrics that are included with the release of each PPMF. And they generate those for the other experimental runs that we do as well.

Now these runs, the ones that are released to the public, are cleared for public release by our

Disclosure Review Board. The other internal runs are not. They are still Title 13 protected until they've gone through DRB review. So, those analyses that are being done on the interim, kind of experimental runs of the DAS, are T13 protected. I don't know if there are plans for a broader release of kind of analysis of some of those runs but yes the demographic directorate has been quite active in reviewing all of the measures that are included in the detail summary metrics and then many more beyond that.

Meghan Maury:
Fantastic. Super helpful. So, wait Michael if I wanted to see those detailed summary metrics that the demographics directorate has created how would I do that?

Michael Hawes:
Absolutely. So again, if you go to our Web site, census.gov, and search “disclosure avoidance,” on the main disclosure avoidance modernizing the 2020 Disclosure Avoidance System Web page there's a link to data metrics, demonstration data metrics. And that is the link to where all of our demonstration data products reside.

The most recent being the April 2021 demonstration data products but also links to all of the earlier versions if you want to do comparisons. And those demonstration data releases contain multiple files. There is the privacy-protected microdata file for the persons universe, the privacy protected microdata file for the units universe and the detailed summary metrics.

Now those detailed summary metrics as I said were generated based on the feedback that we got from many of our data users over the past couple of years on what measures would allow them to most readily assess fitness-for-use of the demonstration data to their specific use cases. So, we've got numerous ways of measuring accuracy. So, we have mean absolute error and mean absolute percent error, mean algebraic percent error. We've got measures of outliers and more. And that's at different levels of geography for different statistics.

We then have specific accuracy measures that are for different types of use cases like total allocation error of shares I think is the acronym, that looks at what the impact on funding decisions might be at various levels of geography and others. So, now the detailed summary metrics do include measures for which there are no data currently.

Many of these error measures were generated for use cases that are supported by the demographic and housing characteristics file, which the demonstration data that we've been releasing the last few cycles have been limited only to the redistricting data. So, you'll see more of those error measures populated when we resume producing demonstration data for eventually for the DHC file but all of the redistricting ones are also included in there.

Meghan Maury:
Awesome. There's a question here about the global epsilon and the epsilon for geographic levels and queries. They're asking if there's a numerical relationship between global epsilon and the epsilon for geographic levels in place. But I wonder if you could just speak to it a little bit more generally - how can people take a look at that allocation and what else they might want to know?

Michael Hawes:
Sure. So, in that same location on our Web site where the detailed summary metrics are there's actually a spreadsheet that gives the exact allocations that were used for the demonstration data runs. But effectively the way you can think of it kind of in the abstract is you assign a value of epsilon to each data products. And those values together added up become your global Privacy-loss Budget.

But you assign a value of epsilon to a particular data product, like the redistricting file. Then, when it's being run through the top down algorithm, you divide that overall Privacy-loss Budget for that data product into slices of the pie, into shares, with a certain share going to everything at the national level, a certain share going to everything at the state level, a certain share going to the county level and so on with all of those shares adding up to 100%.

You similarly do a similar allocation of shares to the different sets of queries that get performed. And then the amount of Privacy-loss Budget that's allocated to any particular query at any particular level of geography is essentially the share for the geographic level multiplied by the share for that particular query.

Now I say essentially because, in reality, one of the major advantages of the top down algorithms design is that tabulations at lower levels of geography inherit accuracy from tabulations that have already been performed at higher levels. So, if you're allocating certain

shares to queries at the national level, queries at the state level are going to inherit some of that accuracy by virtue of the accuracy that's already been determined at the level above and so on. So, it's how much we're allocating to the processing in the noisy measurements at any given level but the actual statistics are going to be more accurate than the allocation specifically.

Meghan Maury:
Yes, that's helpful. And I found it really helpful to look at that table that you mentioned to see sort of where the allocation falls and help me a little bit understand how the whole system works. So, I highly recommend that. We're still getting a couple of questions about. "Is there more time to consider, you know, alternatives within the allocation to address some of the issues that people have raised in their feedback around outliers, et cetera?" Again still getting questions on has DSEP made a final decision? Can we just revisit that one more time?

Michael Hawes:
Sure. So, DSEP is still deliberating as I mentioned before. And as I mentioned at the beginning -- and for those who missed the beginning -- one of the major sets of information that they're reviewing in their deliberations are the extensive and really valuable feedback that we've gotten from our data users on their analyses of the April 2021 demonstration data.

So, thank you to everyone who submitted. They have all been looked at. Every single member of the DSEP has the complete compilation of all of the feedback that we received. And I know they're looking at it closely because I've been getting questions about it from many of them. So, it is all being reviewed and it's being taken very seriously.

As I mentioned, they're still making their deliberations on this. Those deliberations will likely conclude in the very near future. We have a production schedule after all leading us towards that August 16 release date for the redistricting data product.

But again they want to make sure that they do this right. So, they're taking every piece of information they can and making sure that the parameter settings that they make are going to be the right ones. And then those get programmed into the system. We run it. We check it. We check it again. We have our demographic directorate do their through analysis. And then we release the data.

Meghan Maury:
Thank you. Sorry, I apologize for coming back to that but I think it's really important for people to sort of understand where we're at and make sure that they really know what the team is doing. Rolando another question for you. There's a couple questions in here about whether you've conducted additional types of experiments, things that mix suppression and swapping, whether you're planning on releasing any results of these experiments at different levels of geography? Can you talk to us a little bit about how you all are thinking about that if you all are thinking about that?

Rolando Rodriguez:
Well I will say no we have not combined assessing the impact with swapping plus suppression because as I said before we haven't actually implemented suppression. We have assessed how much suppression would be needed. So, we couldn't do that until we've actually implemented a suppression algorithm for 2010 which would be a major effort. I think that would be more effort than we'd be likely to be able to do. We're also going to have to implement one say for 2020.

That being said, you know, if there's great interest in really seeing outcomes associated with swapping experiments that is something we can consider. But as for the main purpose of these have been to inform the budgetary decisions that DSEP is trying to make right now.

Michael Hawes:
And Meghan if I could add on to what Rolando was saying because I think it's a point worth looking at about just how complicated suppression is as a methodology. Rolando touched on the idea that complementary suppressions are necessary in order to protect the data that you're suppressing. So, you suppress the vulnerable cells in the table but then you can, if you don't suppress additional cells, you can recalculate those suppressed values just by subtracting the other reported values from totals that's just a fundamental requirement of suppression methodologies.

But even more challenging in that, and the reason that it would be so difficult and time consuming to implement a suppression strategy is the sheer complexity and interconnectedness of the census publications. The fact that different characteristics are

disaggregated in different ways in different tables. And that all of these tables are nested within a geographically hierarchical structure you have lots of different ways that values could be added up or subtracted.

So, you have to make sure that if you're suppressing a particular value that not only do you have complementary suppressions in that particular table but you also have to verify that you can't recalculate that suppressed value by swapping the reported data from the parent geography's value or that you don't have another table at that level that disaggregates the data in a different way that could be used to subtract out the suppressed value and so on. And going through all of those possible relationships between those tables at all the different levels of geography to find all of your necessary complementary suppressions requires very complicated linear programming to accomplish. And if you miss even one of them the whole thing can topple like a house of cards. So, that's why it's such a complex, such a complex process and why we don't have those results that the question was just asking about.

Meghan Maury:
Yes thanks for that. That's - it's really helpful because at a glance it seems like suppression would be pretty simple, you know, you just say look for values below, you know...

Michael Hawes:
It's really easy to do it wrong...

Meghan Maury:
Yes.

Michael Hawes:
...I'll tell you that. And I've seen lots of agencies at lots of different levels of government who have done their suppression wrong. And I've shown them just how easy it is to make that house of cards tumble down. So, if you do it you want to do it right and that's what takes time and effort.

Meghan Maury: Yes. Thanks for that. Again just one clarifying question -- and I think I know the answer -- but there's one more clarifying question about DSEP. The question was, "Will DSEP have access to the actual letters that were sent in?" And the answer is yes. The

compilation is a compilation of feedback.

Of course there is also some, I'm sure there's some summary information that you provided as well. But I think the questions you were talking about that's DSEP are coming to you with questions on peoples' direct feedback.

Michael Hawes:
They yes every member of DSEP has a copy of every single submission that we received.

Meghan Maury:
Wow that's a lot of reading. There are a couple of questions in here about the re-identification. And there's a sort of a theme between them that I'm seeing which is people are surprised by the fact that these swapping examples, even at a 5% swap, there's so little impact on re-identification that that might seem, you know, not surprising to people but that there was so little impact on reducing re-identification for those high swap rates. Can you talk a little bit about why that is?

Rolando Rodriguez:
So, we haven't assessed the particular swaps to see like where we're not getting re-identification and where we are. I mean I think I will say that, you know, you do get a significant drop when you go to the 50%, just relative to what you get at say 5%, it's just not proportional to the amount of swap that you have done. And that's likely related to just patterns within housing units. And, you know, that's something where we definitely could do a deeper look into it's just numbers we haven't generated yet right now.

Meghan Maury:
Yes that's helpful. I think most of the other questions in here again many of them touch on those legal components and we'll leave those for our colleague attorneys to answer at another time. But I think there are some sort of forward-looking questions too. And I don't know if you are the right people to answer them, but I'll throw them out and see if you feel like you have the information at your fingertips to provide any information.

One of the questions is about the changes to DHC that are coming down the road. I think, you know, as we've been talking to folks about disclosure avoidance and they've been thinking

about how the PL tables are related to the DHC tables, I think folks are curious about why the data that will be released for DHC in 2020 is different from the data that was released back in 2010 under SF1? And I know we don't have time to go into all the details but if you could speak to that a little bit at a general level if you have that information at your fingertips?

Michael Hawes:
Sure. And a member of our demographic directory would be able to answer this much more thoroughly than I can. But there are changes to the publications with every census, tables are added or subtracted. This time around we were more deliberate in our decisions about, like, what tables should be included, largely because when you are using a formally private mechanism of privacy loss accounting, any tabulation that you're producing uses up a share of your Privacy-loss Budget.

And so for any given level of privacy the more tabulations you're producing the less accurate the total will be. So, you want to make sure that when you're expending Privacy-loss Budget because privacy and accuracy are both finite commodities, when you're spending Privacy-loss Budget you want to make sure you're spending it in the most effective and efficient way possible.

So, one of the things our demographic directorate did as they began the planning for the 2020 Census data products was look at the use cases that underlie all of the different tables that have been included. And they tried to essentially strip out of the plan tabulations for which there were no documented use cases. And that was much of what they were doing in terms of thinking about the structure of the 2020 data products.

The other component of that is there are different ways of performing disclosure avoidance and some are better for certain types of data and others methods are better for other types of data. The same holds true for the algorithms that we're using for the 2020 Disclosure Avoidance System. The TopDown Algorithm is really good for producing the formally private microdata that can feed into our tabulation systems, but as we've discussed in other contexts that it also imposes certain constraints on how the algorithm can work.

There are other tabulations, for example the person household joins -- which many of our data users are interested in -- those are much better handled with other formally private algorithms.

And so, some of those, like householder characteristics tables, are being shifted to the later data products that will be protected using this other collection of formally private algorithms rather than being included in the DHC where they would be subjected to the TopDown Algorithm.

Meghan Maury:
Got it. That makes a ton of sense to me. Another sort of forward-looking question and I'm pretty sure that you're not the right person to answer this but I'm going to throw it out anyway just in case. Are we going to provide any kind of training. This person particularly asked about for those folks who are involved in redistricting on how to understand the accuracy metrics and how that might impact their work? They're not necessarily statisticians so they might not be able to digest the metrics quite as easily. How are we helping to make sure that people understand how to use this data if they have anything?

Michael Hawes:
That's a great question. And I would - there's a few pieces there that I can point people to and an action item I can take back to the Census. So, for starters I would say if you're interested in learning how to interpret the accuracy measures that we've included in the detailed summary metrics I would point you to the Webinar that we did a couple of weeks ago with our colleague Matt Spence where he walked through how to understand the different types of accuracy measures in the detailed summary metrics and how to navigate that file. I think that's a really useful starting point.

We also did the Webinar with Tommy Wright and Kyle Irimata where they walked through their analysis of specifically the redistricting use case in the context of the demonstration data. That was a little more technical or depending on who you are a lot more technical and might not be as accessible for somebody without a statistical background.

Separate from that we are working with the Population Reference Bureau, PRB, to develop a handbook that will accompany the redistricting data products. They will explain a lot of what we've done with the Disclosure Avoidance System and how the TopDown Algorithm works and what the implications of differentially private - privacy loss accounting are in this context. So that handbook I think will be a really useful tool for understanding the implications of all of this for the redistricting use case specifically. So, there's a lot of pieces there.

But to the part of the question about we will we be doing more training on it? That's a great question. And it's something that I will certainly consider and I'll bring back and see what we can do.

Meghan Maury:
Yes, and I will say as someone who lives a bit more in the communications world part of what we're hoping to hear from data users over the coming months before that data is released is what kind of training what information would be most useful for folks when we do release that data. I know we've got Jen Shopkorn on the line as well as Shelly who are thinking about these things as well. So we appreciate it as communications folks we appreciate that question.

There are additional questions in here again. I just want to reiterate that DSEP is looking at feedback from everyone who submitted it. There's no piece of feedback that they're not considering to be part of what they need to think about and analyze as they move through the - deliberation process as Michael just mentioned. We have just another moment, Michael, Rolando, I just wanted to give you a second just if there is anything you want us to wrap us up with or anything you wanted to make sure folks heard before we sign off.

Michael Hawes:
I would actually say one thing. We talked about the DHC briefly. And that's something that I know many of our data users care a lot about. In parallel with all of our work to get the redistricting data into production, we are turning our attention towards doing experiments for the demographic and housing characteristics file.

And So, if data users have particular use cases they're interested in there or particularly accuracy targets that they think we should be tuning towards please submit those as well. That input can be very helpful for us as we tune all those little dials and settings that I talked about before. So stay tuned for more on that. And if you have accuracy measures that you want to suggest please submit those for the DHC and we will take them into consideration.

Meghan Maury:
Fantastic, well with that Michael I'm turning it back to you to wrap us up.

Michael Hawes:
Great. Well again thank you everyone for joining us today. This Webinar will be posted on to our Web site if you want to go back to it. Please again sign up for our newsletter to stay informed of all the latest updates. And with that it is now 3 o'clock and I wish you all a great rest of your afternoon and happy weekend.

END